Up and Running with the HathiTrust Research Center

Digital Library Brown Bag Series
November 30, 2016

About Us

- **Nicholae Cline**
  - Scholarly Services Librarian
- **Leanne Nay**
  - Scholarly Technologies Librarian
- **Ewa Zegler-Poleska**
  - IDEASc Fellow

What is the HTRC?

Image from Flickr User Karen Roe

Hottie?

- **Hathi (pronounced hah-tee) is the Hindi word for elephant, an animal highly regarded for its memory, wisdom, and strength.**
  - Trust is a core value of research libraries and one of their greatest assets. In combination, the words convey the key benefits researchers can expect from a first-of-its-kind shared digital repository.

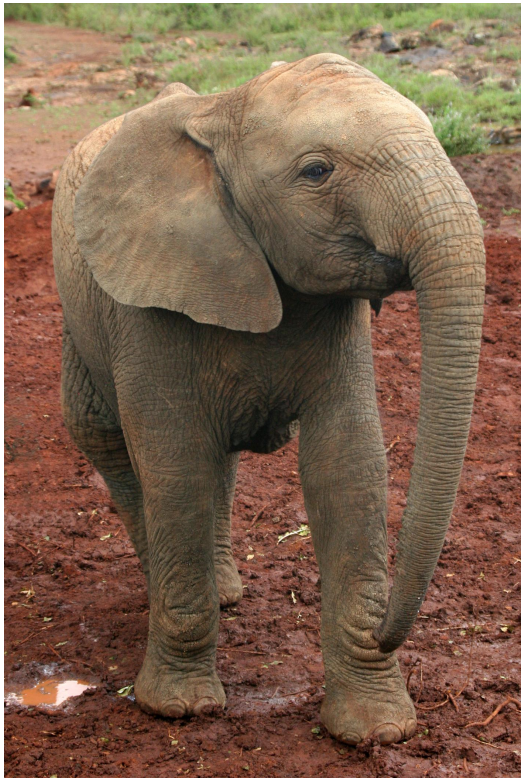
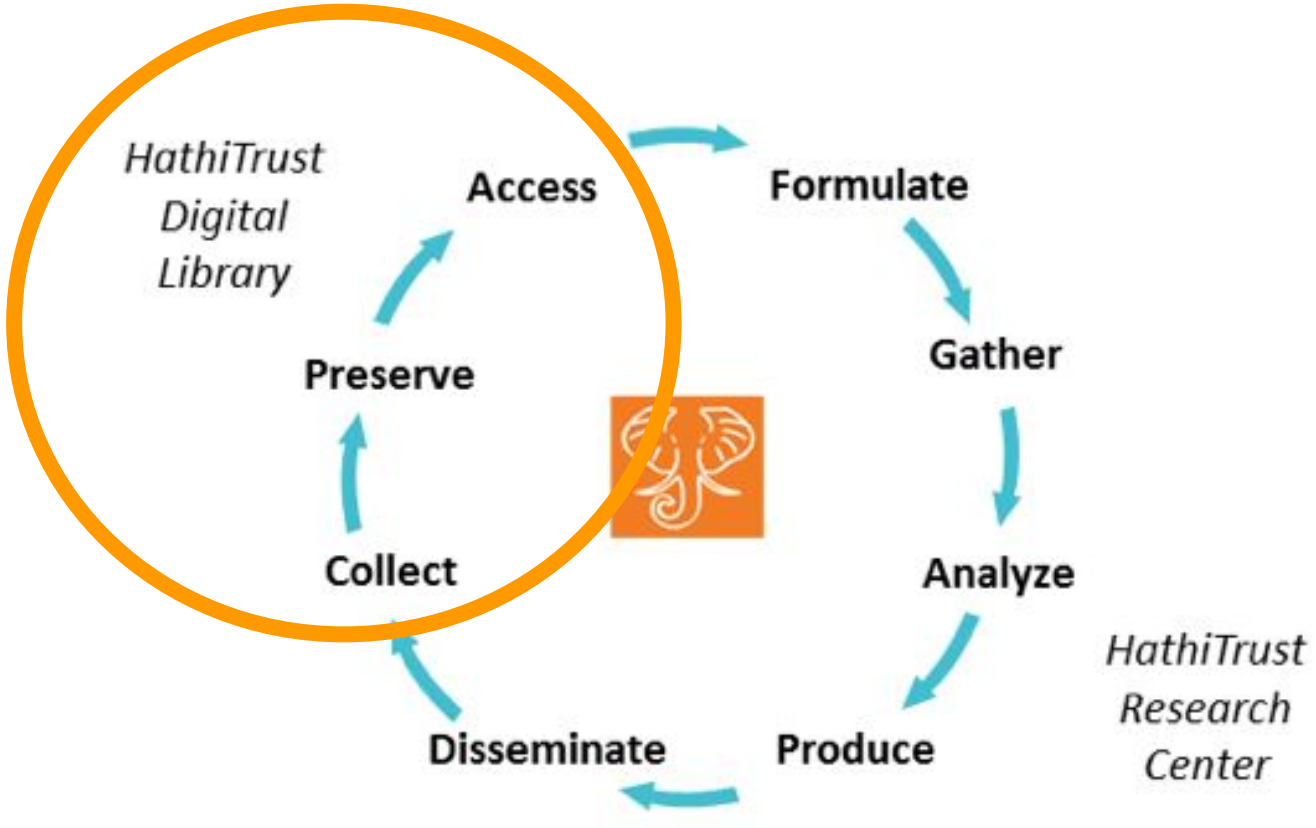


Image from Flickr User Kimberly Brown-Azzarello

About the HathiTrust

- Founded in 2008
- Grew out of large-scale digitization initiative at academic research libraries
  - Google Books project
- 110 member institutions continue to contribute

HathiTrust Ecosystem



HathiTrust Ecosystem

About the HathiTrust Digital Library

- **Repository**
  - 13+ million volumes | 3+ billion pages
  - 50% of volumes are in English
  - Material from the 15th C. on | 20th C. concentration
  - 70% in copyright or undetermined | 30% open
- **Interface**
  - Search and read books in the public domain

HATHI TRUST Digital Library

LOG IN

Search HathiTrust's digital library

FULL-TEXT | CATALOG

Search words about or within the items

Search

Advanced full-text search | Search tips

Full view only

Should I search catalog or full-text?

Want to get the most out of HathiTrust?

Log in with your partner institution account to access the largest number of volumes and features.

Not with a partner institution? »

HathiTrust is a partnership of academic & research institutions, offering a collection of millions of titles digitized from libraries around the world.

WHAT CAN YOU DO WITH HATHITRUST?

BROWSE COLLECTIONS

Explore user-created featured collections.

READ BOOKS ONLINE

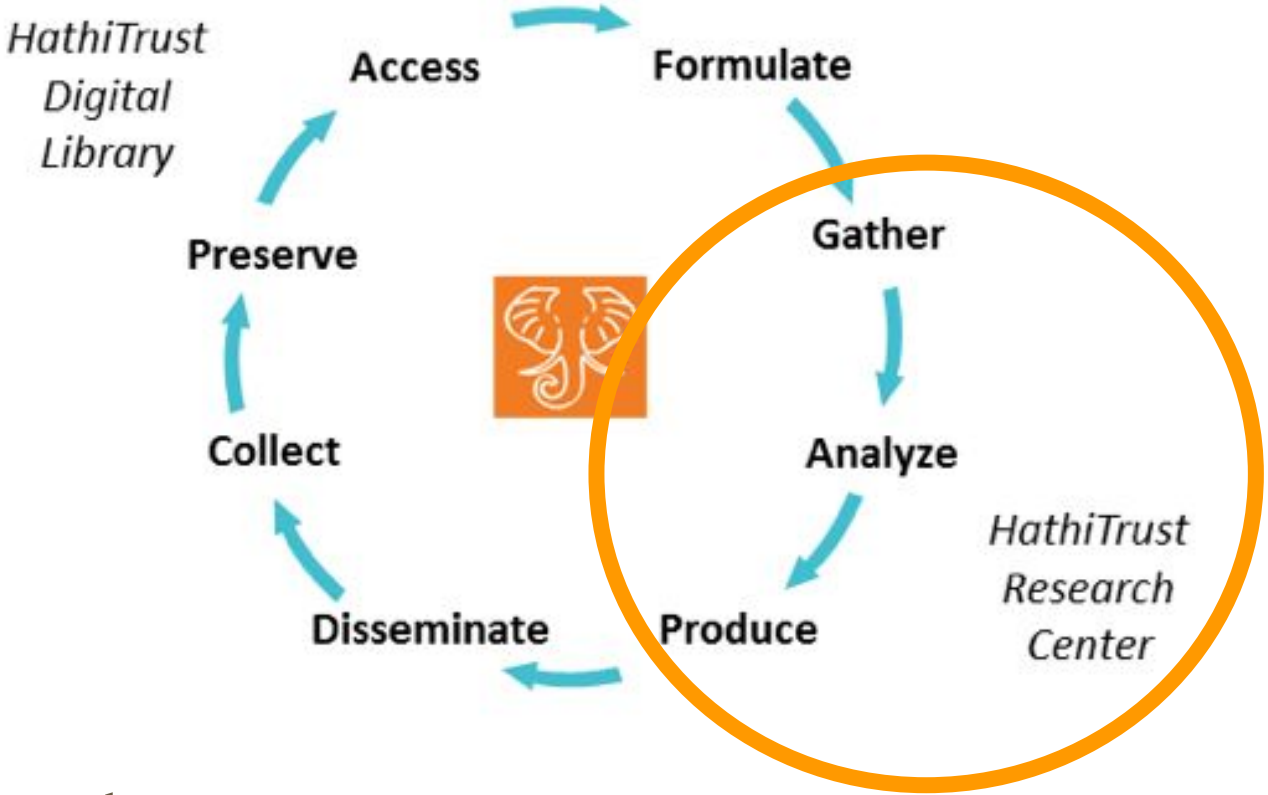
Read millions of titles online — like this one!

READ BOOKS ON THE GO

Take the library's books anywhere with our mobile website.

DOWNLOAD BOOKS* & CREATE COLLECTIONS

**requires institutional login*



HathiTrust Ecosystem

About the HTRC

- Facilitates text analysis of HTDL content
  - Large-scale, computational research
- Research & Development
  - For non-consumptive text analysis
- Located at Indiana University and the University of Illinois

HTRC Tools and Services

- Portal and Workset Builder
- HathiTrust + Bookworm
- HTRC Data Capsule
- Datasets – Extracted Features; Genre in English-language Literature

What is text analysis?

- Text analysis is a subset of data analysis
- Using computers to reveal information in and about text
  - “unstructured” text
  - discern patterns
- What is it used for?
  - identifying spam e-mail
  - detecting plagiarism

What can text analysis do?

- Model aboutness
  - What topics are represented in the text?
- Map connections
  - Where do texts converge?
- Analyze style
  - What does style reveal about authorship?
- Make sense of patterns
  - What features define a group of text?

How does text analysis work?

- Frequency counts
  - Words, phrases, parts of speech, etc
- Collocations
  - Words and phrases that occur close together
- Machine learning
  - Unsupervised (ex. topic modeling)
  - Supervised (ex. classification algorithms)

Text Analysis and Research

- Shift in perspective, shift in research questions
  - Scaling-up
  - More than just search
- Distant reading
  - Franco Moretti
- One step in the research process
  - Can be combined with close reading
  - “Intermediate reading” or “distant-close reading”

Non-Consumptive Research Paradigm

- *Research in which computational analysis is performed on one or more volumes or textual objects in the HTDL, but not research in which a researcher reads or displays substantial portions of an in-copyright or rights-restricted work to understand the expressive content presented within that work.*

- Foundation underlying structure of HTRC work
- Other terms: non-expressive use

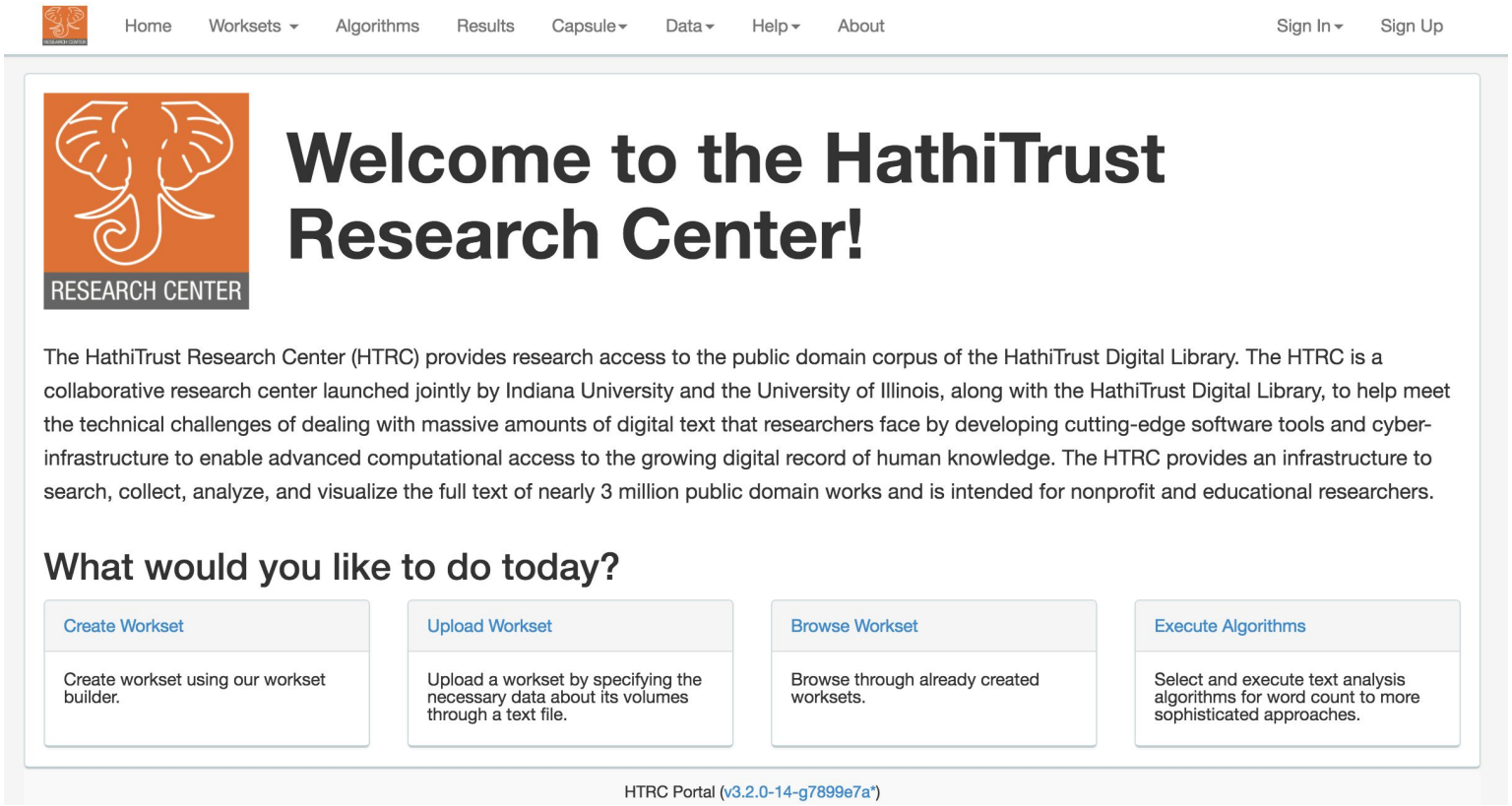
Where do I go from here?

Image from Flickr User Craig Sefton

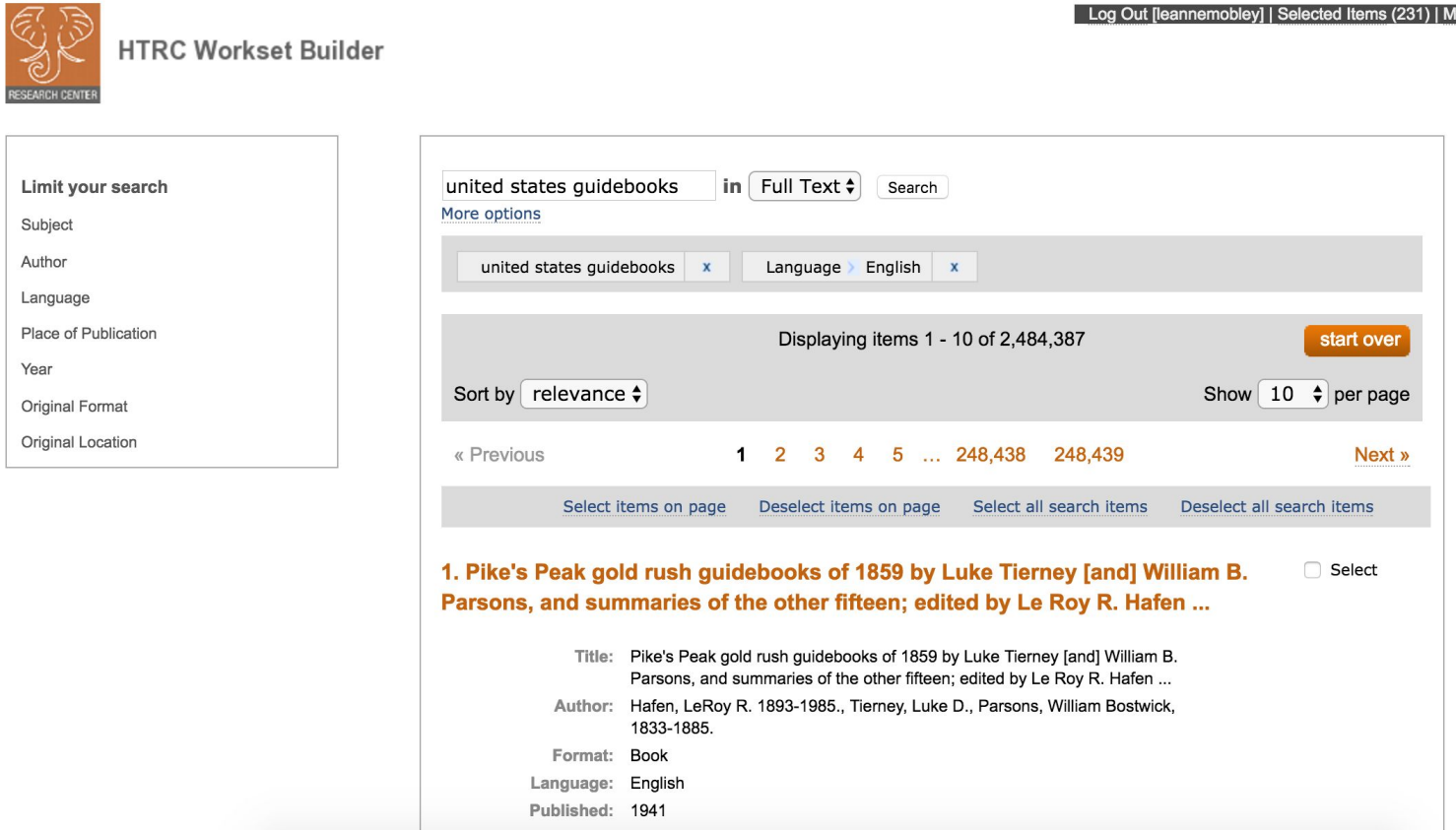
Tools for Beginners

- **Portal and Workset Builder**
  - Build a collection of titles from the HathiTrust Digital Library
  - Run an off-the-shelf algorithm
- **HathiTrust + Bookworm**
  - Visualize language trends over time

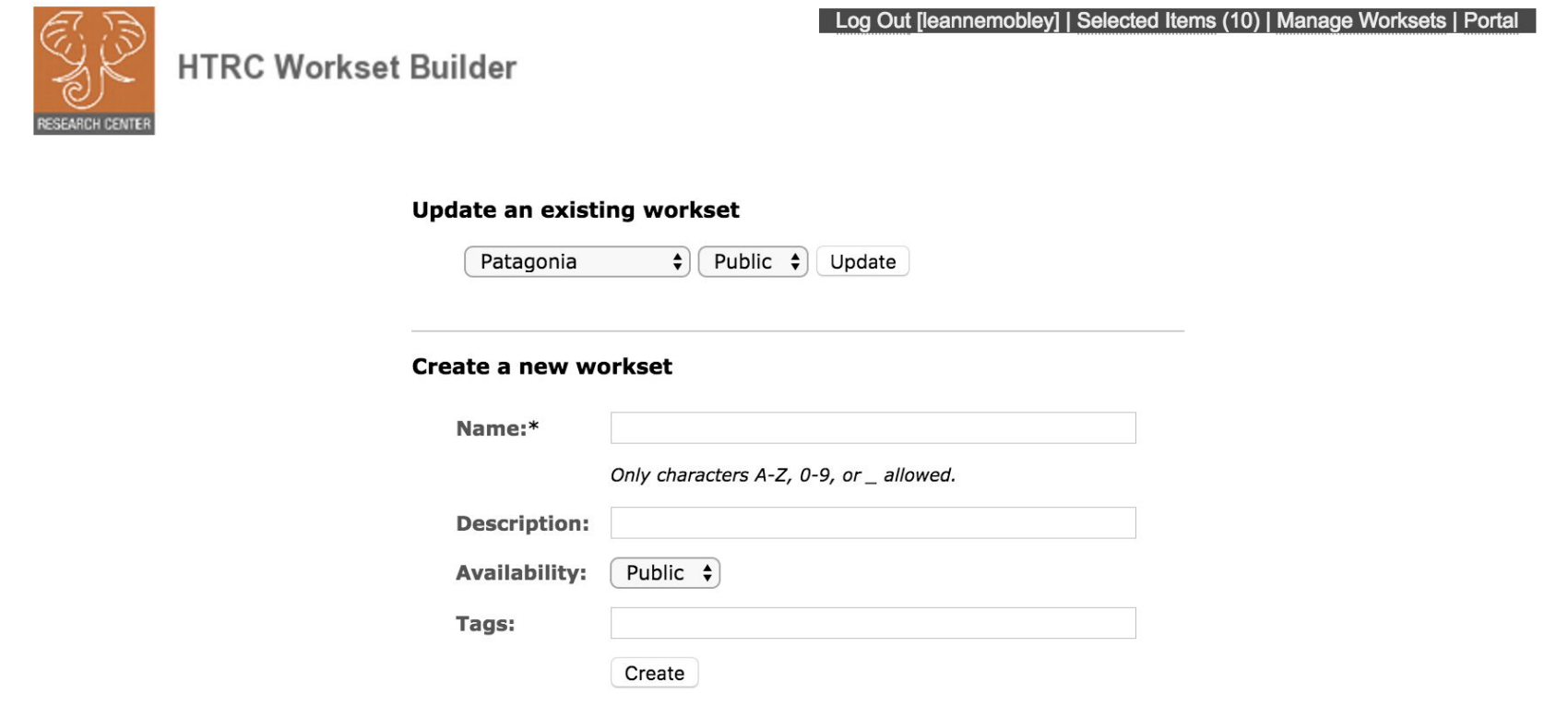
analytics.hathitrust.org


Home
Worksets
Algorithms
Results
Capsule
Data
Help
About
Sign In
Sign Up
RESEARCH CENTER
Welcome to the HathiTrust Research Center!
The HathiTrust Research Center (HTRC) provides research access to the public domain corpus of the HathiTrust Digital Library. The HTRC is a collaborative research center launched jointly by Indiana University and the University of Illinois, along with the HathiTrust Digital Library, to help meet the technical challenges of dealing with massive amounts of digital text that researchers face by developing cutting-edge software tools and cyber-infrastructure to enable advanced computational access to the growing digital record of human knowledge. The HTRC provides an infrastructure to search, collect, analyze, and visualize the full text of nearly 3 million public domain works and is intended for nonprofit and educational researchers.
What would you like to do today?
Create Workset
Create workset using our workset builder.
Upload Workset
Upload a workset by specifying the necessary data about its volumes through a text file.
Browse Workset
Browse through already created worksets.
Execute Algorithms
Select and execute text analysis algorithms for word count to more sophisticated approaches.
HTRC Portal (v3.2.0-14-g7899e7a*)

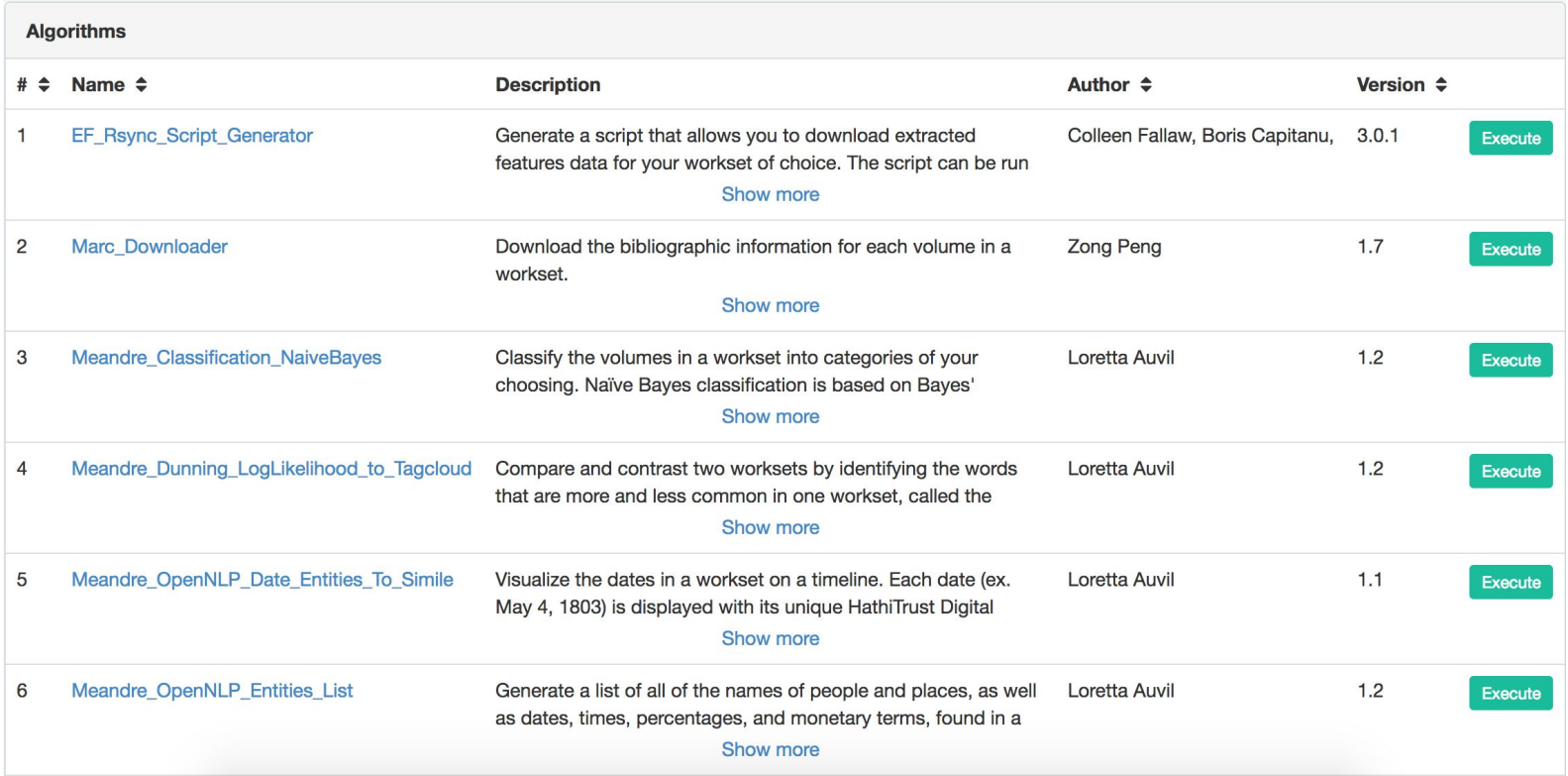
HTRC Portal



Workset Builder



Create a New Workset



Algorithms

#	Name	Description	Author	Version	
1	EF_Rsync_Script_Generator	Generate a script that allows you to download extracted features data for your workset of choice. The script can be run Show more	Colleen Fallaw, Boris Capitanu,	3.0.1	Execute
2	Marc_Downloader	Download the bibliographic information for each volume in a workset. Show more	Zong Peng	1.7	Execute
3	Meandre_Classification_NaiveBayes	Classify the volumes in a workset into categories of your choosing. Naïve Bayes classification is based on Bayes' Show more	Loretta Auvil	1.2	Execute
4	Meandre_Dunning_LogLikelihood_to_Tagcloud	Compare and contrast two worksets by identifying the words that are more and less common in one workset, called the Show more	Loretta Auvil	1.2	Execute
5	Meandre_OpenNLP_Date_Entities_To_Simile	Visualize the dates in a workset on a timeline. Each date (ex. May 4, 1803) is displayed with its unique HathiTrust Digital Show more	Loretta Auvil	1.1	Execute
6	Meandre_OpenNLP_Entities_List	Generate a list of all of the names of people and places, as well as dates, times, percentages, and monetary terms, found in a Show more	Loretta Auvil	1.2	Execute

Algorithms in the Portal

Meandre_OpenNLP_Entities_List

Description: Generate a list of all of the names of people and places, as well as dates, times, percentages, and monetary terms, found in a workset. You can choose which entities you would like to extract.

How it works: Using the OpenNLP system to automatically extract entities:

- loads each page of each volume from HTRC;
- removes the first and last line of each page;
- joins hypenated words that occur at the end of the line;
- extracts entity types specified from the text;
- displays each entity with the volume*id, page*id, sentence_id and character position within the sentence

Note: The volume limit is 100.

Result of job: table of the named entities found in a workset

Version: 1.2
Author: Loretta Auvil
Please enter a job name: (required)

DarwinIndiana@leanne::Meandre_Open::15:40:55 ✔

Named-Entity Recognition Algorithm

Output

named_entities_list.html | stderr.txt | stdout.txt

sentenceId	text	type	textStart	volume_id	page_id
1	Canada	location	43	mdp.39015081775705	8
1	Northern States	location	96	mdp.39015081775705	8
1	Western States	location	135	mdp.39015081775705	8
2	North America	location	139	mdp.39015081775705	8
11	United States	location	131	mdp.39015081775705	8
7	Tenn.	location	56	mdp.39015081775705	9
13	Vermont	location	0	mdp.39015081775705	9
13	Canada	location	12	mdp.39015081775705	9
19	Alexandria	location	3	mdp.39015081775705	9
19	Washington	location	18	mdp.39015081775705	9
25	New York	location	18	mdp.39015081775705	9
28	Boston	location	4	mdp.39015081775705	9
33	Boston	location	3	mdp.39015081775705	9
53	Colorado	location	11	mdp.39015081775705	9
61	Conn.	location	21	mdp.39015081775705	9
62	River	location	0	mdp.39015081775705	9
63	York	location	18	mdp.39015081775705	9
64	Atlanta	location	3	mdp.39015081775705	9

Results

Meandre_Topic_Modeling

Description: Identify "topics" in a workset based on words that have a high probability of occurring close together in the text. Topics are models trained on co-occurring text using Latent Dirichlet Allocation (LDA), where each topic is treated as a generative model and volumes are assigned a probability of how likely each topic is to have generated that text. The most likely words for a topic are displayed as a word cloud.

How it works:

• loads each page of each volume from HTRC;

• removes the first and last line of each page;

• joins hyphenated words that occur at the end of the line;

• removes all tokens that do not consist of alphanumeric characters

• filters stop words;

• replaces "not " with "not_" to deal with negations;

• creates a topic model using Mallet;

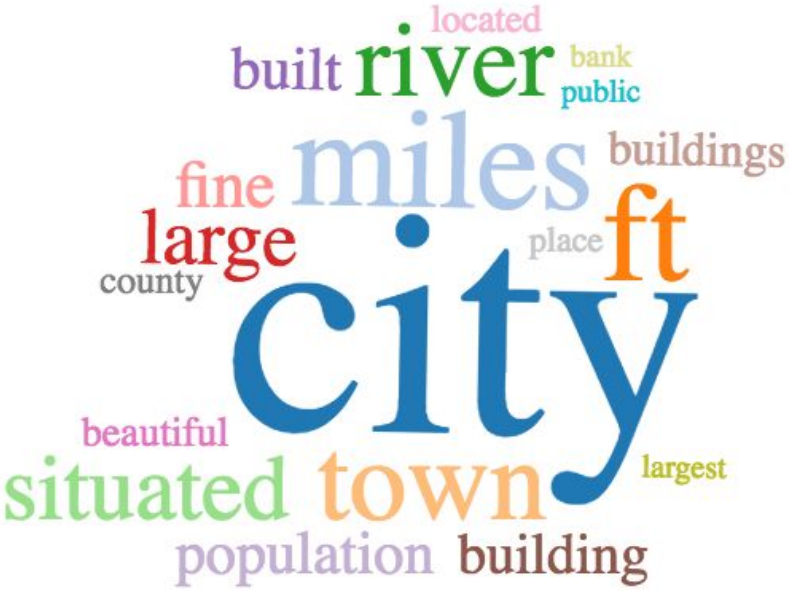
• displays the top 200 tokens in a tag cloud

Note: The upper limit on the number of volumes is 1000.

Result of job: xml file with topics, and visualizations of them in the form of tag clouds.

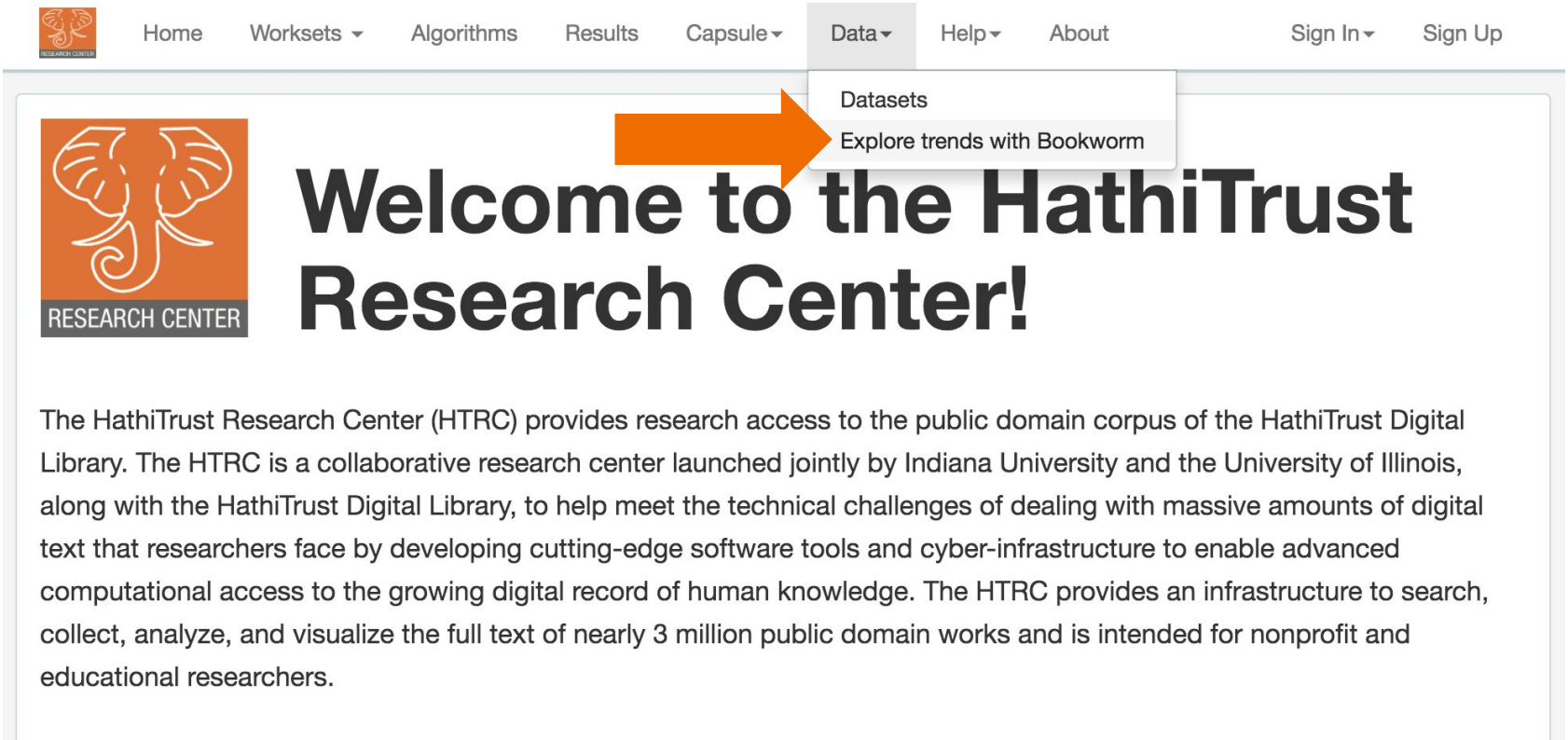
Version: 1.2
Author: Loretta Auvil

Topic Modeling Algorithm

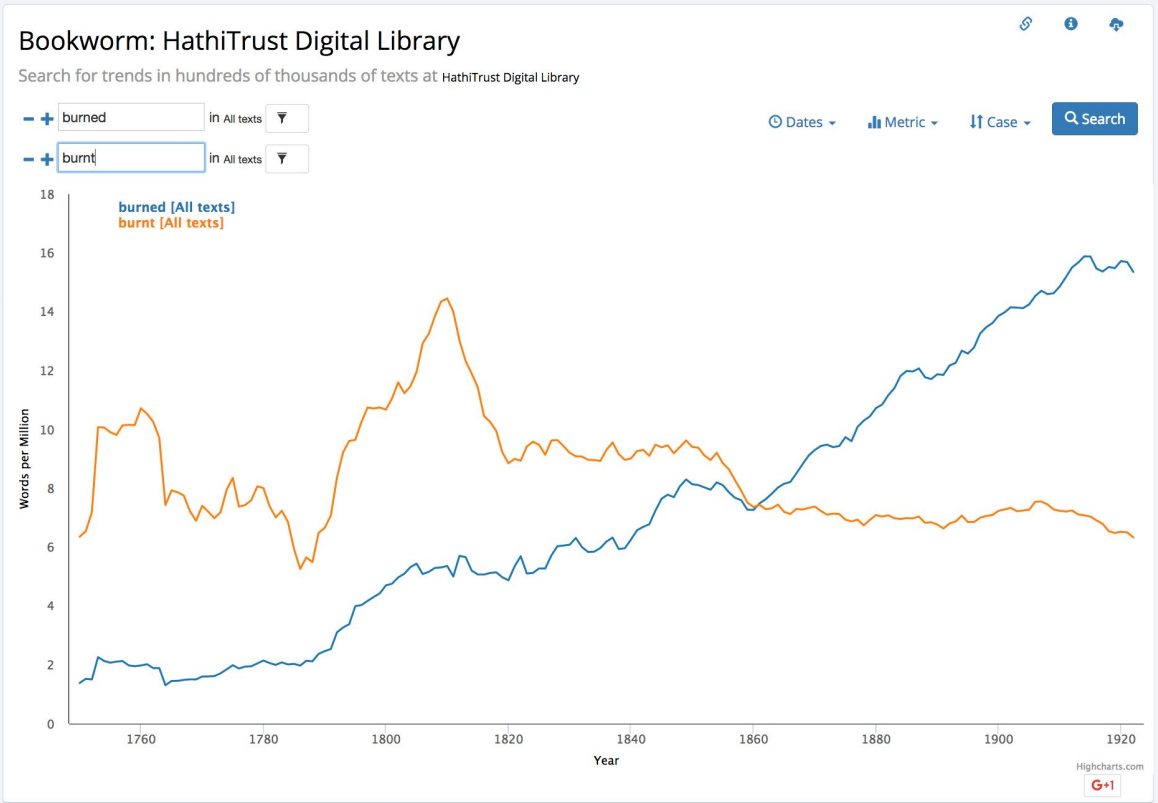

located
built
river
bank
public
miles
buildings
fine
large
place
ft
county
city
beautiful
largest
situated
town
population
building


turn
fork
left
right-hand
poles
end
bridge
bear
straight
follow
page
road
macadam
trolley
4-corners
ahead
iron
left-hand
crossing

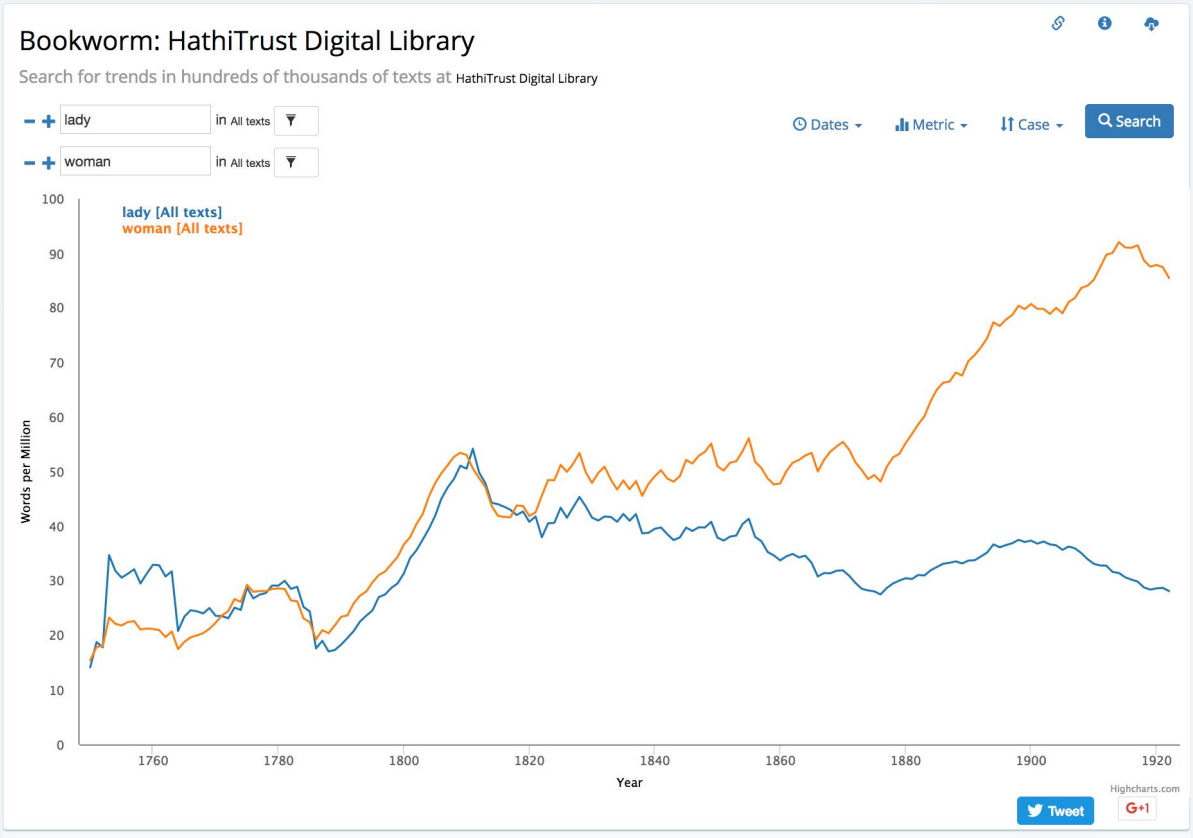
Results



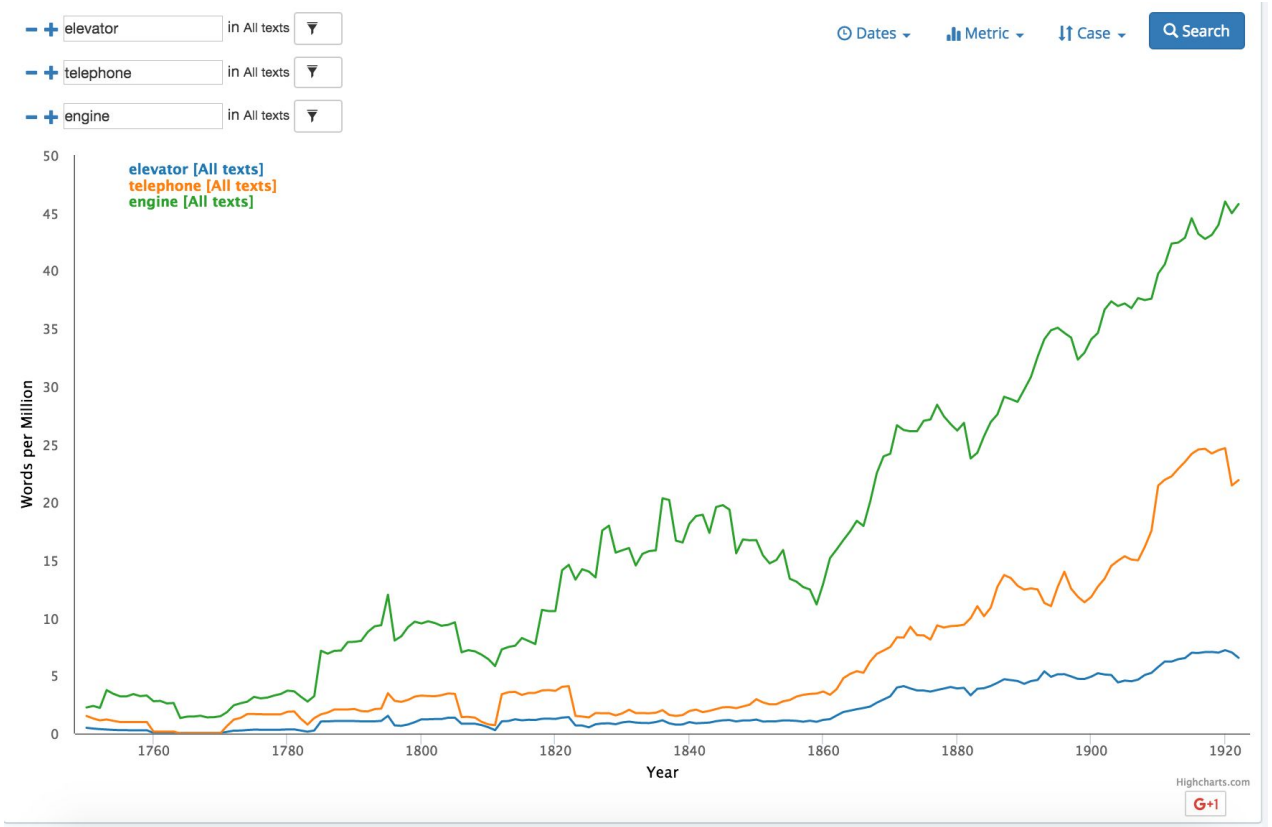
HathiTrust + Bookworm



‘Burned’ vs. ‘Burnt’ over time



‘Lady’ vs. ‘Woman’ over time



‘Elevator’ ‘Telephone’ and ‘Engine’ over time

How do we fit in?

Image from Flickr User Adair Broughton

Scholarly Commons Working Group

- **Education & Outreach**
  - IMLS grant “Digging Deeper, Reaching Further”
  - Scholars’ Commons consultation hours
  - Host workshops and give presentations at other institutions
  - Conduct user interviews
    - Transcribe and code responses

Digging Deeper, Reaching Further

- **IMLS funded project**
- **PI**
  - Harriett Green, University of Illinois at Urbana-Champaign Library
- **Co-PIs**
  - Angela Courtney, Indiana University-Bloomington Libraries
  - J. Stephen Downie, University of Illinois
  - Neil McElroy, Lafayette College Library
  - Geoffrey Morse, Northwestern University Libraries
  - Beth Sandore Namachchivaya, University of Illinois at Urbana-Champaign Library
  - Amanda Henley, University of North Carolina at Chapel Hill Library

DDRF Project Goals

Through development of curricular materials to be disseminated through a Train the Trainer program, we will:

- **Provide librarians with new content** for instructional services that address the curricular and research needs of students and faculty around digital scholarship and digital humanities;
- **Empower librarians** to become active research partners on digital projects at their institutions;
- Provide the foundation to **transform academic libraries' scholarly commons and digital humanities centers into more data-intensive collaborative learning spaces**, both physically and virtually, through use of this curriculum and engagement in community dialogues on digital humanities resources.

DDRF Progress

- **Spring 2016**
  - Pilot workshops at University of Illinois and Indiana University
- **Fall 2016**
  - Pilot workshops at partner institutions
  - Lafayette College, Northwestern University, University of North Carolina

DDRF 2017

- Continue reworking module content based on feedback from workshop participants and facilitators at partner institutions
- Host another series of pilot workshops

IDEASc Fellowship

- Integrated Doctoral Education with Application to Scholarly Communication
- Fellowship program designed to further scholarship and practice in the area of scholarly communication by integrating practical experience in the library with the traditional doctoral research and classroom experiences
- Funding provided by IMLS
- Started in 2016

IDEASc IMLS Grant Team

Cassidy Sugimoto

Associate Professor, School of Informatics and Computing, Indiana University Bloomington

Julie Bobay

Associate Dean for Collection Development and Scholarly Communications, Indiana University Libraries, Indiana University Bloomington

Elin K. Jacob

Associate Professor, Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington

Carolyn Walters

Executive Associate Dean of the IU Libraries and Executive Director of the Office of Scholarly Publishing, Indiana University Bloomington

John A. Walsh

Associate Professor, Department of Information and Library Science, School of Informatics and Computing, Indiana University Bloomington

Nazareth Pantaloni, III

Copyright Program Librarian, Indiana University Libraries, Indiana University Bloomington

IDEASc Fellows

- Information and Library Science PhD students
  - Pei-Ying Chen
  - Shawn Martin
  - Jennifer St. Germain
  - Ewa Zegler-Poleska

- Library placements - semester/year-long positions
  - Copyright, data management, digital projects, OA publishing...
  - Digital Collections, ScholarWorks, MDPI, HTRC...

Advanced Collaborative Support Grants

- RFP issued about once a year
- Award for dedicated HTRC developer support and/or time
- First round was in Spring 2015
  - Next call should be Spring 2017
- Projects require access at scale, using half a million to a million volumes each

Get Involved

HTRC Announcements:

htrc-announce-l@list.indiana.edu

HTRC User Group:

htrc-usergroup-l@list.indiana.edu

Thank you!

Image from Flickr User Russ Allison Loar